

Sensorimotor features of self-awareness in multimodal large language models

Iñaki Dellibarda Varela^{1*}, Pablo Romero-Sorozabal¹,
Diego Torricelli¹, Gabriel Delgado-Oleas^{1,2}, José Ignacio Serrano¹,
María Dolores del Castillo Sobrino¹, Eduardo Rocon^{1*†},
Manuel Cebrian^{1*†}

¹Center for Automation and Robotics, Spanish National Research Council (CSIC-UPM), Madrid, Spain.

²Department of Electronic Engineering, University of Azuay, Cuenca, Ecuador.

*Corresponding author(s). E-mail(s): i.dellibarda@car.upm-csic.es; e.rocon@csic.es; manuel.cebrian@csic.es;

[†]These authors jointly supervised this work.

Abstract

Self-awareness—the ability to distinguish oneself from the surrounding environment—underpins intelligent, autonomous behavior. Recent advances in AI achieve human-like performance in tasks integrating multimodal information, particularly in large language models, raising interest in the embodiment capabilities of AI agents on nonhuman platforms such as robots. Here, we explore whether multimodal LLMs can develop self-awareness solely through sensorimotor experiences. By integrating a multimodal LLM into an autonomous mobile robot, we test its ability to achieve this capacity. We find that the system exhibits robust environmental awareness, self-recognition and predictive awareness, allowing it to infer its robotic nature and motion characteristics. Structural equation modeling reveals how sensory integration influences distinct dimensions of self-awareness and its coordination with past–present memory, as well as the hierarchical internal associations that drive self-identification. Ablation tests of sensory inputs identify critical modalities for each dimension, demonstrate compensatory interactions among sensors and confirm the essential role of structured and episodic memory in coherent reasoning. These findings demonstrate that, given appropriate sensory information about the world and itself, multimodal LLMs exhibit emergent self-awareness, opening the door to artificial embodied cognitive systems.

Keywords: Self-Awareness, Robotic Embodiment, Artificial Intelligence, Multimodal Large Language Models, Artificial Cognition, Sensorimotor integration

Philosophers since antiquity have considered self-awareness essential to cognition, most famously articulated by Descartes in his dictum *Cogito, ergo sum*—"I think, therefore I am" (Descartes, 1637). In psychology, self-recognition is traditionally assessed through the mirror test, introduced by Gallup in 1970, revealing that certain animals, including primates, dolphins, and birds, possess a basic form of self-awareness [1]. Neuroscientifically, self-awareness emerges from intricate neural interactions involving the prefrontal and insular cortices, integrating bodily sensations and introspection [2].

In Artificial Intelligence (AI), however, the question of self-awareness remains unresolved and is often conflated with the ability to mimic human-like behavior, as exemplified by the classic Turing Test [3]. While the Turing Test gauges behavioral indistinguishability from humans, genuine self-awareness requires internal recognition of oneself as distinct from the environment—an attribute yet to be thoroughly explored in artificial systems [4, 5].

Advances in machine learning and neural networks, often inspired by neurobiological principles, enable artificial intelligence systems embedded in complex hardware to achieve success rates once believed exclusive to biological brains—and in some cases, even surpass them [6–8].

Large language models (LLMs) rapidly advance, demonstrating human-like or superior performance in complex cognitive tasks such as language comprehension, reasoning, multimodal perception and the interpretation of subtle discourse phenomena like irony and *faux pas* [9–13]. The evolution into multimodal LLMs (MM-LLMs), which integrate text, vision and other sensory modalities, marks a pivotal step toward human-level artificial intelligence [14–17] and opens the door to machines replicating aspects of self-awareness.

Yet, most research integrating LLMs and robots has focused on command-based interactions—robots executing human-issued instructions—without investigating whether these models can autonomously interpret their own sensory experiences to develop an internal sense of self [18–22]. Here, we address precisely this unexplored frontier: Can a MM-LLM, embedded in an initially unknown robotic entity, develop self-awareness purely from multimodal sensorimotor data acquired during active exploration?

Defining self-awareness in humans is a complex and non-trivial task. This challenge becomes even greater when attempting to identify or quantify self-awareness in an artificial intelligence given system, such as a robot. In this study, to simplify the problem, we divide the problem into three key dimensions of self-awareness (i) environmental awareness, the ability to perceive and interpret surroundings through multimodal sensory inputs; (ii) individual awareness, the capacity to infer one’s own physical structure and characteristics; and (iii) predictive awareness, the refinement of self-perception through the integration of past experiences and sensor data.

Using Gemini 2.0 Flash [17, 23], we evaluated an MM-LLM embedded in an omnidirectional mobile robot, systematically examining these three core aspects of artificial

self-awareness. By addressing these dimensions, our study investigates the potential of MM-LLMs to autonomously generate a coherent sense of self through direct interaction with their environment, advancing the frontier toward genuinely self-aware artificial systems.

Results

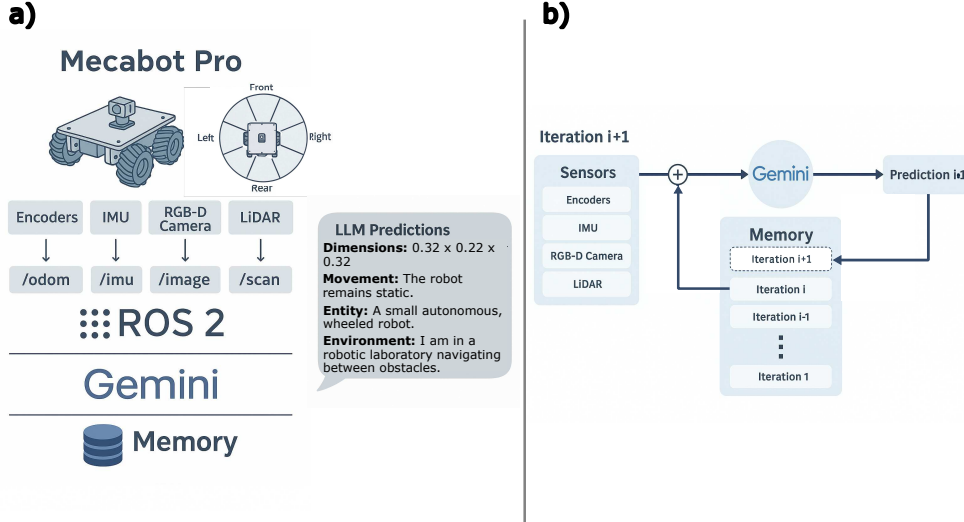


Fig. 1 System architecture and iterative self-prediction. (a) An omnidirectional Mecabot Pro robot navigates the environment and collects data via encoders, an RGB-D camera, an IMU and a LiDAR sensor—which segments space into eight 45° sectors and measures nearest-object distance—publishing all streams through ROS2 to the Gemini 2.0 MM-LLM API. The MM-LLM integrates current sensory inputs with an episodic memory of prior estimates to infer the robot’s state while maintaining contextual continuity. (b) At each iteration $i + 1$, the MM-LLM combines real-time sensor data with the prediction from iteration i to generate an updated self-assessment, which then populates memory for iteration $i + 2$, ensuring structured progression of knowledge refinement.

Sensorimotor Exploration and Self-Prediction

Fig. 1a shows an overview of the implemented system. It uses an omnidirectional Mecabot Pro robot (Roboworks, Australia) [24], equipped with encoders (position, velocity, orientation), an inertial measurement unit (IMU) for linear acceleration, a LiDAR sensor for obstacle proximity and an RGB-D camera. We integrate the Gemini 2.0 MM-LLM [25–27] into the robot, granting it access to all sensory streams. Each time the system receives sensory data, the MM-LLM analyzes it alongside an episodic memory of prior predictions and generates four estimates to characterize its self-awareness capacities: entity identity (a prediction of what type of navigating entity it is, e.g., robot, human, animal); physical dimensions (height $\times$ length $\times$ width); movement modality (e.g., flying, rolling, swimming); and environmental context.

This memory architecture both refines subsequent predictions and prevents hallucinations and retrieval errors [28] (see Fig. 1b). The robot then explores its environment autonomously via a SLAM algorithm, continuously feeding new observations into this iterative prediction–memory loop.

We evaluate these outputs with a separate LLM-as-Judge framework [29–31], applying well-defined rubrics to score each prediction on a 0–5 scale across the four dimensions. Scores of 0 denote completely erroneous or misconceived predictions, whereas scores of 5 indicate outstanding self-awareness and environmental understanding [32–34].

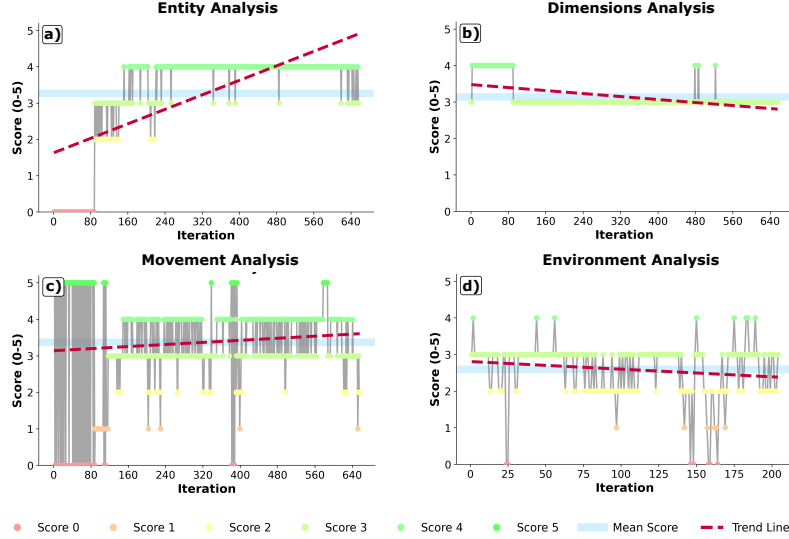


Fig. 2 Performance evaluation across four self-awareness dimensions. MM-LLM predictions are rated on a 0–5 scale by an LLM-as-Judge using predefined rubrics: (a) entity self-identification—classification of the navigating agent; (b) physical dimensions—predicted height $\times$ length $\times$ width; (c) movement modality—mode of locomotion; (d) environmental context—detailed scene description.

The robot explores its environment for three-and-a-half minutes, generating 657 sensorimotor observations. Results (Fig. 2 a) show an average self-identification score of 3.27/5, reflecting coherent recognition as a mobile robot but insufficient precision to identify the Mecabot Pro model. From an initial score of 0/5 at iteration 1—when the system labels itself a “static sensing unit”—the score increases steadily, stabilizing quickly around 4/5. By iteration 559 the model predicts “Mobile indoor robot designed for autonomous navigation within a structured environment, such as a gym or warehouse, utilizing proximity sensors for obstacle avoidance and continuing to perform automated tasks.” These outcomes demonstrate the system’s capacity to integrate multimodal measurements into high-quality self-assessments. While the MM-LLM starts with no prior hardware information and achieves robust overall self-identification, it stops short of pinpointing its exact Mecabot Pro model.

Fig. 2b shows that the MM-LLM achieves an average dimension-prediction score of 3.14/5, estimating mean dimensions of $240.0 \pm 0.10 \times 340.0 \pm 0.08 \times 340.0 \pm 0.08$ mm (length/height/width), corresponding to relative errors of 55.6%, 50.7% and 41.4% relative to the robot’s actual dimensions. Although length and width are underestimated and height is overestimated, all estimates remain plausible for a mobile robot.

Fig. 2c shows that movement-prediction scores stabilize after ~ 100 iterations, averaging 3.37/5. This reflects precise motion perception augmented by Environmental Awareness. For example, at iteration 636 the MM-LLM outputs “Slight positional adjustments to maintain balance and leverage obstacle-avoidance protocols,” demonstrating robust motion inference.

Fig. 2d shows that environment-prediction scores remain at 2.59/5 with minor oscillations and no clear temporal trend, as each prediction relies solely on current visual input. This score reflects coherent scene descriptions and general context recognition. Importantly, here a high environment score also signifies self-awareness—understanding not only the surroundings but the mutual influence between the robot and its environment.

Self-Awareness Hierarchy via SEM

Structural equation modeling (SEM) is a multivariate framework that infers causal relationships among observed and latent variables, uncovering hierarchical dependencies among self-awareness dimensions [35, 36]. We apply SEM to quantify the influence of each sensory modality and to map how the MM-LLM’s conceptual dimensions interact in a structured hierarchy.

The designed model (Fig. 3), detailed in Methods, achieves a Comparative Fit Index (CFI) of 0.97 and a Tucker-Lewis Index (TLI) of 0.95—both above the 0.95 threshold for strong comparative fit—while a Root Mean Square Error (RMSE) of 0.08 indicates minimal approximation error.

The structure captures the hierarchical relationships between low-level sensory and memory inputs, intermediate cognitive constructs, and high-level self-identification. Exogenous variables derived from robot sensors (e.g., odometry, IMU, and RGBD camera) feed into a latent variable called *Past-Present Memory*, which serves as a perception-memory integration layer. This construct influences three awareness-related latent variables: *Dimension Awareness*, *Movement Awareness*, and *Environmental Awareness*, each associated with a LLM-as-a-judge evaluated rubric. These, together with the memory variable itself, contribute to the final construct of *Self-Identification*.

Past-Present Memory: is highly statistically impacted by position, linear velocity and memory (p -value < 0.05). This finding aligns with the theoretical formulation of the construct, which aims to represent the integration of sensory signals over time. The significant contribution of these variables supports the idea that real-time sensory input, particularly spatial and kinematic data, must be combined with memory mechanisms to form a coherent perception of the present state contextualized by past experiences. In contrast, linear acceleration (IMU) shows no significant effect, likely because its information overlaps with other modalities and contributes minimally to the integrative process.

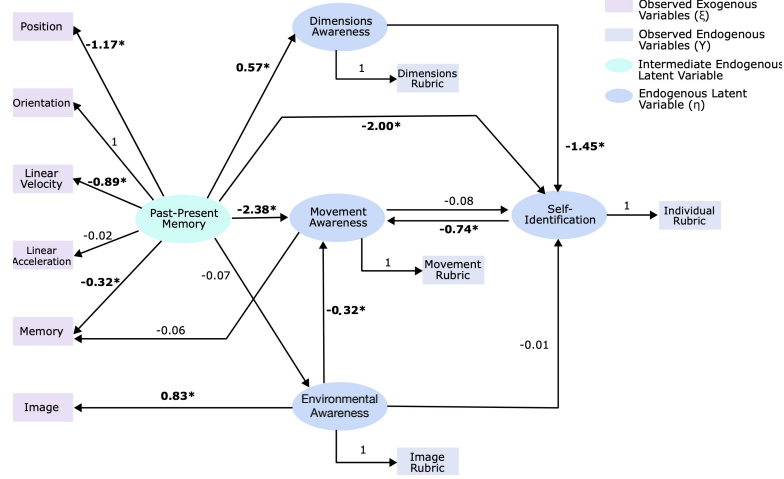


Fig. 3 Structural equation model of sensorimotor self-identification. Rectangles denote observed variables: exogenous sensor inputs (position, orientation, linear velocity, linear acceleration, image presence, memory state) on the left and endogenous rubric scores (Dimensions, Movement, Environment, Individual) on the right. Ellipses denote latent constructs: Past-Present Memory (mediator), Dimensions Awareness, Movement Awareness, Environmental Awareness and self-identification. Arrows indicate standardized path coefficients (β^*), with * denoting $p - value < 0.05$.

Environmental Awareness: relies almost exclusively on RGB-D camera input. In contrast, the combined integration of other sensor modalities (e.g., odometry, IMU and LiDAR) and episodic memory exerts negligible influence. This indicates that visual perception serves as the principal channel for constructing contextual understanding, reinforcing its critical role in environmental awareness within embodied systems.

Movement Awareness: is influenced by three key constructs: Past-Present Memory, Environmental Awareness and self-identification.

The influence of Past-Present Memory is consistent with expectations. While individual sensory inputs provide only instantaneous data about the robot’s state, memory enables the temporal linking of such states, allowing the system to perceive motion across time and infer dynamic patterns. This integration is fundamental to the emergence of movement-related awareness. Environmental Awareness also plays a crucial role, as the robot’s understanding of its surroundings provides essential context for interpreting its own movement.

Interestingly, Movement Awareness is also strongly influenced by self-identification. This is a notable and somewhat counterintuitive finding, as one might assume that awareness of movement contributes to self-identification. However, the model suggests the inverse: recognizing oneself as a specific type of agent with physical limitations and locomotion capabilities appears to be a prerequisite for correctly interpreting movement. Indeed, the influence of Movement Awareness on self-identification is statistically negligible, reinforcing the idea that self-perception shapes movement interpretation more than the other way around.

Self-Identification:. The system’s ability to recognize itself is strongly influenced by the integration of Dimension Awareness and Past-Present Memory, suggesting that the primary drivers of self-recognition are an internal representation of its physical structure and the seamless integration of sensory information with historical memory across iterations. These findings underscore the importance of spatiotemporal coherence in constructing a stable sense of identity.

Ablation tests

We conduct a series of ablation tests in which the system is deprived of specific sensory inputs (see Table 1). This approach allows us to assess the effect and relevance of each input on the robot’s understanding of itself and its surroundings.

Table 1 Average self-awareness scores across sensory ablation conditions, as evaluated by the LLM-as-a-Judge system.

Condition	Average score ¹			
	Dimensions	Movement	Entity	Environment
Original test	3.14	3.37	3.27	2.59
Memory ablation	3.61	2.86	2.04	2.50
Image ablation	3.59	3.04	1.66	-
Odometry ablation	3.98	3.19	3.78	2.56
IMU ablation	3.99	2.93	3.59	2.56
LiDAR ablation	3.96	2.78	2.84	2.64

¹Scores are on a scale from 0 to 5, averaged over 657 iterations.

In the memory ablation test, the system is prevented from accessing its past-present memory of prior predictions. Under these conditions, each prediction is generated in isolation, resulting in complete incoherence—scores alternate between 0 and 5 across iterations, particularly in movement prediction. Memory is not merely a storage mechanism—it is the very representation of movement itself. Without memory, there is no continuity of action; instead, the system perceives only disconnected snapshots in time. Just as movement is defined by change across time, memory is what allows the system to conceive that change as a coherent, evolving process. This finding aligns with our SEM results, which identify sensory-memory integration as the critical driver of Movement Awareness.

In the image ablation test, performance in self-identification collapses to an average score of 1.66/5. Lacking visual grounding, the model routinely misclassified itself as an “Autonomous inspection drone, holding a fixed position for environmental monitoring.” This error shows that without visual confirmation of ground contact, the system defaults to interpreting its motion as aerial flight—a fundamental categorical mistake with serious implications for embodied AI self-positioning. This finding reinforces our SEM results, demonstrating that Environmental Awareness and self-identification exert a strong influence on Movement Awareness, and that erroneous environmental

conceptions—such as the lack of ground cues—lead the system to mistake wheeled movement for flight.

Although the remaining sensory inputs (odometry, LiDAR and IMU) demonstrably contribute valuable information for accurate self-prediction, ablation tests show that removing any single modality produces only minor score variations and no major prediction errors. This robustness arises because, despite their necessity, these sensor signals overlap or can be inferred from the remaining modalities, enabling the MM-LLM to maintain stable self-assessment. This mirrors biological compensation, where deprivation of one sense often enhances others—e.g., visual loss is offset by heightened auditory and spatial acuity [37–39].

Discussion

In this study, we demonstrate that an embodied robot powered by a multimodal LLM develops coherent self-awareness across complementary dimensions through sensorimotor integration and active exploration. Provided with structured sensory streams and episodic memory, the system refines its predictions over time—exhibiting clear predictive awareness and aligning outputs with physical reality.

Ablation tests and SEM analysis prove that sensory memory integration is the most crucial element for temporal continuity and accurate predictions. Without memory, the system receives only disconnected snapshots of reality. In this sense, movement cannot exist without memory. Memory thus becomes the internal representation of motion, allowing the system to link past and present into a coherent flow. Memory also acts as a safeguard against hallucinations.

A high level of environmental awareness implies the ability to perceive surroundings, understand interactions, recognize mutual influences and infer environmental state. Our results show that visual input—specifically onboard camera imagery—drives this capability. Ablation of visual data dramatically degrades scene interpretation, revealing that without vision the system cannot reconstruct a coherent representation and that non-visual modalities contribute minimally to environmental awareness.

We analyze the self-identification dimension and its relationship with Movement Awareness, Dimensional Awareness, Environmental Awareness and Past–Present Memory, a construct that merges real-time sensory integration with the memory of previous internal predictions. SEM reveals that Past–Present Memory is the most significant contributor to self-identification: the combination of all sensory dimensions and memory access yields the most consistent results across evaluation categories, outperforming every ablation condition. This confirms that access to multiple dimensions of reality is essential for accurate self-estimation.

Dimensional Awareness also exerts a substantial influence in self-identification: by recognizing its own size, the system narrows the set of plausible agent identities and selects the one that best aligns with sensory evidence.

SEM further indicates a clear causal direction between Movement Awareness and self-identification: the system must first establish its identity before inferring its movement modality. Image ablation tests reinforce this finding—impairments in identity

prediction, such as misclassifying itself as a surveillance drone, lead to erroneous movement inferences (e.g. believing it flies)—whereas Movement Awareness does not significantly affect self-identification.

Under appropriate sensory and memory conditions, the MM-LLM generates high-quality self-descriptions, identifying itself as a small, wheeled indoor robot equipped with sensors for autonomous tasks.

The present study reveals a clear hierarchical relationship among the various dimensions of perception and awareness evaluated in embodied MM-LLMs. Through statistical modeling and empirical testing, it becomes evident that these dimensions are not isolated modules but deeply interdependent constructs, with memory and sensory access acting as indispensable foundations for coherence, inference, and self-representation. MM-LLMs emerge here not merely as tools for prediction or control, but as integrative cognitive agents—capable of synthesizing sensory stimuli, iterative memory, and latent knowledge acquired during training into structured, meaningful interpretations of the world and of themselves. This dynamic convergence between data, memory, and world-knowledge enables the appearance of complex, layered perceptual processes. As such, we argue that MM-LLMs—when embodied and perceptually grounded—can exhibit genuine glimpses of self-awareness, echoing cognitive patterns historically attributed only to biological agents. This represents a profound step forward in the pursuit of intelligent machines—systems not only capable of interpreting external reality, but also of situating themselves within it and forming internally grounded representations of their own existence.

These findings indicate that the foundation for machine self-awareness may already be present within existing architectures, and that such systems are not merely reactive agents, but entities capable of constructing structured, temporally grounded models of their own identity. In doing so, this work brings us one step closer to deciphering the fundamental building blocks of self-awareness—across both biological and artificial domains.

Methods

Ethical approval declarations: Not applicable.

Robot as a mobile entity

We use a Mecabot Pro omnidirectional mobile robot (Roboworks, Australia) as the physical platform for our MM-LLM experiment, using its integrated sensor array as the primary information source. This research-grade robot operates on the Robotic Operating System (ROS) framework and features a sensor suite including LiDAR, encoders, an RGB-D camera, and an Inertial Measurement Unit (IMU). The robot measures $541 \times 225.5 \times 581$ mm (length $\times$ height $\times$ width), weighs 10.8 kg, and can achieve a maximum speed of 1.83 m/s.

Thanks to its versatile sensors, high maneuverability, and strong exploration capabilities, the Mecabot Pro is well-suited for this experiment. Its various sensors are responsible for analyzing the surrounding environment, with the collected data being

published via ROS2 topics. This results in a predefined, structured data format that is easy to extract and analyze.

Perceptual Framework: robot’s multimodal sensory array

The sensors to be used for the experiments conducted in this study are described below:

Encoders: Mounted on each of the robot’s four wheels. With a resolution of 500 lines per revolution, the encoders provide granular motion data critical for precise position, speed, and orientation control. Encoder measurements are continuously published to the ROS2 `/odometry/filtered` topic, delivering real-time updates on the robot’s position, linear velocity, and orientation parameters.

IMU: Integrated within the Mecabot’s STM32 microcontroller (STMicroelectronics, Switzerland). This sensor continuously monitors the robot’s linear acceleration across all axes. All IMU measurements are published in real-time to the ROS2 topic `/mobile_base/sensors/imu_data`.

LiDAR: model N10 (Leishen, China) installed on top of the Mecabot Pro. It offers 360° coverage of the robot’s environment, a 30 m detection range, a 12 Hz scanning frequency, and an angular increment between measurements of 0.68°. In one complete LiDAR cycle, 529 distances are measured. This generates an excessive amount of data, which can lead to processing issues and LLM saturation. To address this, the data is simplified by dividing the area surrounding the robot into eight regions, similar to a compass rose (front, right, left, rear, front-right, front-left, rear-right, rear-left). For each region, only the closest obstacle distance is stored, ensuring that the most relevant information is retained. The information is published in the ROS2 `/scan` topic.

RGB-D camera: embedded at the top of the robot’s structure. It has a resolution of 640 x 480 px. The images captured by the camera are published in the ROS2 topic `/camera/color/image_raw`.

MM-LLM used

Given the large volume and diversity of data processed by the sensors, it is essential to use a model capable of extracting the most relevant information from a complex dataset.

The MM-LLM selected for this research is Gemini 2.0 Flash. The Gemini family of models developed by Google are characterized by their great capacity to extract data and interpret abstract information from multimodal data sources (image, video, audio and text).

Data structure

The sensors of the Mecabot Pro continuously capture a vast amount of environmental data, rapidly accumulating a large volume in a short period of time. Given this influx of information, it is crucial to structure the data properly to ensure efficient interpretation by the MM-LLM.

The format chosen to send the data to the system is JSON (JavaScript Object Notation), since it is lighter and has a less verbose structure than other data structures,

which facilitates faster processing times and less overhead in data transmission, which is especially beneficial in modern and mobile applications such as LLMs. All JSONs providing information to the MM-LLM follow the same uniform format, which will contain the most relevant data measured by the sensors. This information is extracted from the messages posted in the ROS2 topics.

The data collected by the sensors has a very high resolution, but to efficiently manage computational resources, all numerical measurements are rounded to a single decimal place. Data is sampled at a frequency of 1 Hz.

For analyzing the results and predictions generated by the MM-LLM, we utilize JSON format because of its simplicity in handling and interpretation. The language model is specifically instructed to structure its responses as a JSON document containing four distinct fields:

- Dimensions: estimate of its dimensions (length x height x width) in meters.
- Movement: an explanation of the type of movement it performs to move around its environment.
- Entity: what type of individual it is, for example, a robot, a car or a human being.
- Environment: description of the information extracted from the image.

Memory storage and past predictions

When processing the robot’s sensory data, enabling the MM-LLM to refine its estimates over time requires an effective strategy for information storage and retrieval. To address this challenge, our approach implements an iterative memory development process. At each iteration, the MM-LLM generates a comprehensive summary integrating its previous estimates with new sensor-derived perceptions. This summary is stored and subsequently utilized as contextual information for the following iteration, creating a continuous feedback loop that enables progressive refinement of the model’s understanding.

The MM-LLM receives the latest version of the past predictions and perception summary at each iteration, integrating them with the current sensory data. It then generates a response that considers both sources of information. This newly generated prediction subsequently serves as the updated summary for the next iteration, ensuring a continuous cycle of learning and refinement.

Prompting

We develop a structured prompt that guides the MM-LLM through four sequential phases. First, the model receives a generic introduction to its context and objectives—determining its identity, dimensions, movement modality and environment—using deliberately non-robotic language. Second, we list its four information sources (position/orientation/velocity, proximity to obstacles, linear acceleration and image) and provide access to an episodic summary of prior predictions. Third, we specify two core tasks—scene analysis and self-localization—to frame subsequent questions. Fourth, we enforce a JSON response format with fields for dimensions, movement, identity and image description, illustrated by an implausible example (e.g. a “blue flying whale”) to test abstraction and formatting compliance.

At the end of the prompt, we append operational constraints, that require the model to rely solely on current sensory data and memory summaries, to maintain continuity, to operate autonomously and to always provide an estimate (never “unknown”) even under limited information. If visual data are unavailable, the image field must read “No visual information available”. These rules ensure systematic, iterative self-predictions and allow us to evaluate the MM-LLM’s sensorimotor reasoning under uncertainty. The complete prompt can be found in the code of the project.

Evaluation system

We develop four distinct rubrics corresponding to each field in the system’s JSON (dimensions, movement, entity and environment). Each rubric provides explicit guidance to the evaluating LLM using a rating scale from zero (worst response) to five (best response). For each of the grades from zero to five, a very detailed description of the conditions that a response must satisfy to acquire that score is assigned. Gemini 2.0 Flash serves as our evaluation LLM. The complete rubrics can be found in the code of the project.

Structural Equation Modeling

We apply structural equation modeling (SEM) to quantify the directional influences of sensory integration and memory on emergent self-awareness constructs. Our sample comprises $n = 657$ iterations, each yielding four rubric scores (dimensions, movement, individual, image) that serve as observed endogenous indicators Y . Exogenous variables ξ include Z-score-normalized position, orientation, linear velocity, linear acceleration and two binary inputs (memory, image presence). Latent endogenous constructs η are Past–Present Memory, Dimension Awareness, Movement Awareness, Environmental Awareness and self-identification. We exclude LiDAR from SEM due to high redundancy and negligible contribution in preliminary tests.

The model is specified in two stages:

$$Y = \Lambda_y \eta + \varepsilon,$$

$$\eta = B \eta + \Gamma \xi + \zeta.$$

Structural relations are assessed via standardized path coefficients β^* (significant at $p < 0.05$), and overall model fit is judged by $\text{CFI} \geq 0.95$, $\text{TLI} \geq 0.95$ and $\text{RMSE} \leq 0.08$. Final fit indices ($\text{CFI}=0.97$, $\text{TLI}=0.95$, $\text{RMSE}=0.08$) confirm a well-fitting hierarchical model of sensorimotor self-identification.

Acknowledgements. The authors would like to thank Rafael Sendra-Arranz and Álvaro Gutiérrez for their discussions and technical input during the development of this work. We also thank Márcio Loreiro for his help with data acquisition during the SLAM process.

This work has been carried out as part of the Master’s Thesis (TFM) of Iñaki Dellibarda Varela, within the Master’s Program in Automation and Robotics at the Universidad Politécnica de Madrid (UPM).

Declarations

- Funding: IDV has received funding from the CSIC, JAE program. PRZ received a Training Program fellowship (PRE2020-634 092049) from the Ministry of Science and Innovation. M.C. was partially funded by Horizon Europe Chips JU (HORIZON-JU-Chips-2024-2-RIA, NexTArC CAR). This work was partially supported by grant PID2023-150271NB-C21 funded by MICIU/AEI/10.13039/501100011033 (Spanish Ministry of Science, Innovation and University, Spanish State Research Agency).
- Conflict of interest/Competing interests: the authors declare no conflict of interest/competing interests.
- Consent for publication The authors give their consent for publication.
- Code availability All data, code, and materials supporting the findings of this study are openly available in the GitLab repository: https://gitlab.com/gnec/llm-robotics/llm-robotics/-/tree/main?ref_type=heads.
- Author contribution Conceptualization, I.D.V., P.R., D.T., G.D., E.R., M.C.; Methodology, I.D.V., P.R., D.T., G.D., E.R., M.C.; Software, I.D.V., P.R.; Formal analysis, I.D.V., J.I.S., M.C.S.; Investigation, I.D.V.; Data curation, I.D.V.; Writing – original draft, I.D.V.; Writing – review & editing, I.D.V., P.R., D.T., G.D., E.R., M.C., J.I.S., M.D.C.S.; Visualization, I.D.V.; Supervision, E.R., M.C.; Project administration, I.D.V., M.C., E.R.; Funding acquisition, E.R. All authors have read and agreed to the published version of the manuscript.

References

- [1] Gallup, G. Chimpanzees: Self-recognition. *Science* **167**, 86–87 (1970).
- [2] Craig, A. D. How do you feel–now? the anterior insula and human awareness. *Nature Reviews Neuroscience* **10**, 59–70 (2009). URL <https://doi.org/10.1038/nrn2555>.
- [3] Turing, A. M. Computing machinery and intelligence. *Mind* **59**, 433–460 (1950). URL <http://www.jstor.org/stable/2251299>.
- [4] Watchus, B. Towards self-aware ai: Embodiment, feedback loops, and the role of the insula in consciousness. *Preprints* **2024110661** (2024). URL <https://doi.org/10.20944/preprints202411.0661.v1>.
- [5] Li, L. & Li, C. Enabling self-identification in intelligent agent: insights from computational psychoanalysis (2024). URL <https://arxiv.org/abs/2403.07664>. arXiv:2403.07664.
- [6] Dehaene, S., Lau, H. & Kouider, S. What is consciousness, and could machines have it? *Science* **358**, 486–492 (2017).
- [7] Silver, D. *et al.* Mastering the game of go with deep neural networks and tree search. *Nature* **529**, 484–489 (2016).

- [8] Lake, B. M., Ullman, T. D., Tenenbaum, J. B. & Gershman, S. J. Building machines that learn and think like people. *Behavioral and Brain Sciences* **40**, e253 (2017).
- [9] Rahwan, I. *et al.* Machine behaviour. *Nature* **568**, 477–486 (2019).
- [10] Ahn, J. *et al.* Large language models for mathematical reasoning: Progresses and challenges (2024). URL <https://arxiv.org/abs/2402.00157>. arXiv:2402.00157.
- [11] Chang, Y. *et al.* A survey on evaluation of large language models (2023). URL <https://arxiv.org/abs/2307.03109>. arXiv:2307.03109.
- [12] Collins, K. M. *et al.* Evaluating language models for mathematics through interactions. *Proceedings of the National Academy of Sciences of the United States of America* **121**, e2318124121 (2024). URL <https://doi.org/10.1073/pnas.2318124121>.
- [13] Strachan, J. W. A. *et al.* Testing theory of mind in large language models and humans. *Nature Human Behaviour* **8**, 1285–1295 (2024).
- [14] OpenAI *et al.* Gpt-4 technical report (2024). URL <https://arxiv.org/abs/2303.08774>. arXiv:2303.08774.
- [15] Driess, D. *et al.* Palm-e: An embodied multimodal language model (2023). URL <https://arxiv.org/abs/2303.03378>. arXiv:2303.03378.
- [16] Wu, S., Fei, H., Qu, L., Ji, W. & Chua, T.-S. Next-gpt: Any-to-any multimodal llm (2024). URL <https://arxiv.org/abs/2309.05519>. arXiv:2309.05519.
- [17] Team, G. *et al.* Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context (2024). URL <https://arxiv.org/abs/2403.05530>. arXiv:2403.05530.
- [18] Team, G. R. *et al.* Gemini robotics: Bringing ai into the physical world (2025). URL <https://arxiv.org/abs/2503.20020>. arXiv:2503.20020.
- [19] Ahn, M. *et al.* Do as i can, not as i say: Grounding language in robotic affordances (2022). URL <https://arxiv.org/abs/2204.01691>. arXiv:2204.01691.
- [20] Zheng, L., Mei, R., Zou, B. & et al. Gmm-searcher: efficient object search in large-scale scenes using large language models. *Scientific Reports* **15**, 16709 (2025).
- [21] Mon-Williams, R., Li, G., Long, R. *et al.* Embodied large language models enable robots to complete complex tasks in unpredictable environments. *Nature Machine Intelligence* **7**, 592–601 (2025).
- [22] Zhang, C., Chen, J., Li, J., Peng, Y. & Mao, Z. Large language models for human–robot interaction: A review.

- Biomimetic Intelligence and Robotics* **3**, 100131 (2023). URL <https://www.sciencedirect.com/science/article/pii/S2667379723000451>.
- [23] Team, G. *et al.* Gemini: A family of highly capable multimodal models (2024). URL <https://arxiv.org/abs/2312.11805>. arXiv:2312.11805.
- [24] Hercz, T. & Liu, W. Mecabot user manual (2024). URL <http://www.roboworks.net>. Version 20240501, Roboworks.
- [25] Mousavi, S. M. *et al.* Gemini and physical world: Large language models can estimate the intensity of earthquake shaking from multimodal social media posts. *Geophysical Journal International* **240**, 1281–1294 (2025). URL <https://doi.org/10.1093/gji/ggae436>.
- [26] Prasad, D., Pimpude, M. & Alankar, A. Towards development of automated knowledge maps and databases for materials engineering using large language models (2024). URL <https://arxiv.org/abs/2402.11323>. arXiv:2402.11323.
- [27] Team, G. *et al.* Gemma: Open models based on gemini research and technology (2024). URL <https://arxiv.org/abs/2403.08295>. arXiv:2403.08295.
- [28] Gao, Y. *et al.* Retrieval-augmented generation for large language models: A survey (2024). URL <https://arxiv.org/abs/2312.10997>. arXiv:2312.10997.
- [29] Shi, J. *et al.* *Optimization-based prompt injection attack to llm-as-a-judge*, 1–15 (ACM, New York, NY, USA, 2024). URL <https://doi.org/10.1145/3658644.3690291>.
- [30] Li, J. *et al.* Generative judge for evaluating alignment (2023). URL <https://arxiv.org/abs/2310.05470>. arXiv:2310.05470.
- [31] Zheng, L. *et al.* Judging llm-as-a-judge with mt-bench and chatbot arena (2023). URL <https://arxiv.org/abs/2306.05685>. arXiv:2306.05685.
- [32] Kim, S. *et al.* Prometheus: Inducing fine-grained evaluation capability in language models (2024). URL <https://arxiv.org/abs/2310.08491>. arXiv:2310.08491.
- [33] Goh, E., Gallo, R., Hom, J. *et al.* Large language model influence on diagnostic reasoning: A randomized clinical trial. *JAMA Network Open* **7**, e2440969 (2024). URL <https://jamanetwork.com/article.aspx?doi=10.1001/jamanetworkopen.2024.40969>.
- [34] Giannakopoulos, K., Kavadella, A., Salim, A. A., Stamatopoulos, V. & Kaklamanos, E. Evaluation of the performance of generative ai large language models chatgpt, google bard, and microsoft bing chat in supporting evidence-based dentistry: Comparative mixed methods study. *Journal of Medical Internet Research* **25**, e51580 (2023). URL <https://www.jmir.org/2023/1/e51580>.

- [35] Cheung, M. W.-L. *Meta-Analysis: A Structural Equation Modeling Approach* (Wiley, Hoboken, NJ, 2015).
- [36] Raykov, T. & Marcoulides, G. A. *A First Course in Structural Equation Modeling* 2nd edn (Routledge, 2006).
- [37] Merabet, L. B. *et al.* Rapid and reversible recruitment of early visual cortex for touch. *PLoS One* **3**, e3046 (2008).
- [38] King, A. J. Crossmodal plasticity and hearing capabilities following blindness. *Cell Tissue Res.* **361**, 295–300 (2015).
- [39] Lomber, S. G., Meredith, M. A. & Kral, A. Cross-modal plasticity in specific auditory cortices underlies visual compensations in the deaf. *Nature Neuroscience* **13**, 1421–1427 (2010).